Raja Gond

Microsoft Research India | IIT Bombay'23

rajagond.github.io raja.gond@outlook.com github.com/rajagond Google Scholar

EDUCATION

Indian Institute of Technology (IIT) Bombay, Mumbai, MH, India

- Bachelor of Technology (B.Tech) in Computer Science and Engineering — Jul 2019 – May 2023
  - with Honors.
  - Cumulative GPA: 8.6 / 10.0

PUBLICATIONS

PREPRINTS

[2] Raja Gond, Nipun Kwatra, and Ramachandran Ramjee, "TokenWeave: Efficient Compute-Communication Overlap for Distributed LLM Inference," in *Arxiv.DC*, May 2025.
DOI: 10.48550/arXiv.2505.11329
Code: https://github.com/microsoft/tokenweave (26 stars) (In Submission)
Presented as a poster at the *Microsoft Research India Academic Summit*, June 24, 2025.

[1] Raja Gond and Purushottam Kulkarni, "emucxl: an emulation framework for CXL-based disaggregated memory applications," in *Arxiv.DC*, Apr 2024.
DOI: 10.48550/arXiv.2404.08311
Code: https://github.com/cloudarxiv/emucxl (15 stars)

RESEARCH EXPERIENCE

Microsoft Research Lab, Bengaluru, KA, India

- Pre-doctoral Researcher, AI-Infrastructure — Jul 2023 – Present
  - Project: Compute-Communication Overlap for Efficient Distributed LLM Inference
  - Advisors: Dr. Nipun Kwatra and Dr. Ramachandran Ramjee
  - **TokenWeave:**
    - Co-authored *TokenWeave*, a system for efficient compute–communication overlap in distributed LLM inference, achieving up to $1.29\times$ latency and $1.26\times$ throughput improvements on $8\times$H100 GPUs over the vLLM baseline, outperforming state-of-the-art solutions such as TileLink and NanoFlow
    - Designed and implemented a fused **AllReduce–Residual–RMSNorm** kernel using NVIDIA Hopper's NVSHARP and Multimem features, reducing GPU SM usage to just 2–8 SMs and enabling compute–communication overlap in vLLM, with performance gains even at small batch sizes and short sequence lengths
  - **Before TokenWeave – Overlap for Efficient Inference in Mixture-of-Experts (MoE) Models:**
    - Implemented *Expert Parallelism in vLLM* and demonstrated its benefits Tensor Parallelism for MoE models
    - Designed a lightweight signaling mechanism to initiate Direct Memory Access (DMA)-based partial GPU–GPU communication, freeing all SMs for compute and enabling effective compute–communication overlap
    - Achieved up to a **20%** reduction in MoE MLP latency for Mixtral-22B in microbenchmarks on $8\times$H100 GPUs
  - **Other Contributions:**
    - Conducted an in-depth analysis of all prior compute–communication overlap techniques (e.g., TileLink, Flux, NanoFlow), identifying limitations in their applicability to modern GPU architectures and emerging models

Dept. of Computer Science and Engineering, IIT Bombay, Mumbai, MH, India

- Undergraduate Researcher, SynerG Lab — Aug 2022 – Jun 2023
  - Project: emucxl: Emulation Framework and Access Library for CXL-Based Disaggregated Memory Systems
  - Advisor: Prof. Purushottam Kulkarni
    - Developed a user-space library coupled with a **NUMA-based CXL emulation backend** for standardized CXL memory access that enables rapid prototyping of disaggregated memory solutions
    - Conducted a literature survey on CXL standards and showed emucxl capabilities through practical use cases
  - Project: Persistent Memory (PMem) Applications [PDF, code]
  - Advisors: Prof. Purushottam Kulkarni and Prof. Umesh Bellur
    - Designed and implemented a robust reader-writer program on Non-Volatile Memory using advanced array and pointer techniques, which provides fault tolerance and efficient data access
    - Explored **Persistent Memory Development Kit** libraries to understand PMem capabilities and analyzed performance differences between traditional and PMem-based Redis using real-world benchmarks

INDUSTRY EXPERIENCE

Morgan Stanley, Mumbai, MH, India

- Technology Analyst Intern, Investment Management Division — May 2022 – Jul 2022
  - Designed and implemented a Java utility library for translating MT Swift payment messages generated by a trading platform into enriched MX messages, facilitating and streamlining the migration process to new messaging standards
  - Integrated MX format verification and conducted in-depth analysis of MT formats, MX equivalents, and translation
  - Received an offer for a **full-time position** with the team upon graduation, based on exemplary internship performance

TEACHING

Undergraduate Teaching Assistant, Dept. of Computer Science and Engineering, IIT Bombay

- Computer Networks + Lab (CS224/CS252) — Spring'23
  - Instructor: Prof. Bhaskaran Raman
  - Responsible for evaluating lab assignments, explaining concepts, and resolving doubts for over **200 CSE sophomores**
- Operating Systems + Lab (CS347/CS333) — Fall'22
  - Instructors: Prof. Purushottam Kulkarni and Prof. Umesh Bellur
  - Designed and managed lab assignments, addressed student doubts during lab sessions and online, proctored theory and lab exams, and evaluated answer scripts and lab coding assignments, for a batch of over **180 CSE juniors**
- Computer Systems (Bootcamp) — Summer'22
  - Instructors: Prof. Purushottam Kulkarni and Prof. Mythili Vutukuru
  - Involved in the design of weekly assignments and asynchronous doubt-solving to aid self-paced learning for students

MENTORSHIP

Department Academic Mentor, Student Mentorship Program, IIT Bombay — Jul 2022 – Apr 2023

- Selected out of **70+** applicants through a rigorous procedure based on SoP, interviews, and peer reviews
- Mentored students with academic or general concerns to help ease their transition into the CSE department

SERVICE

Artifact Evaluation Committee: OSDI/ATC'25, SOSP'25

AWARDS & SCHOLARSHIPS

- Microsoft Global Hackathon 2023: Executive Challenge First Prize Award — Sep 2023
  Hack for the Microsoft Cloud in the Era of AI (Idea: Microsoft Confidential)
  Collaborated closely with the Hackathon teammates spread across global Microsoft offices to develop an innovative solution that enhances cloud infrastructure capabilities and presented it to the Microsoft Cloud + AI leadership
- Research Fellowship, Microsoft Research India — Jul 2023 – Jul 2025
  Selected as **one of 30** Research Fellows at **Microsoft Research India** from a pool of **12,000+** applicants
- Merit-cum-Means Scholarship, IIT Bombay — Jul 2019 – May 2023
  Awarded the Merit-cum-Means Scholarship during undergraduate studies

SELECTED ACADEMIC PROJECTS

Dept. of Computer Science and Engineering, IIT Bombay

- SCLP: Compiler for C-like Language — Spring'22
  Guide: Prof. Uday Khedker — Implementation of Programming Languages
  - Built a compiler to generate Abstract Syntax Tree (AST), Three Address Code, and corresponding assembly Code
  - Implemented the scanner using **Lex**, the parser using **Yacc** and constructed the object-oriented AST representation in **C++**, enabling the efficient processing of arithmetic and relational expressions, loops, and control flow statements
- Custom Shell and Feature Extension of xv6 — Fall'21
  Guide: Prof. Mythili Vutukuru — Operating Systems
  - Implemented custom shell supporting serial, parallel, and background command execution with signal handling
  - Designed and implemented a **priority-based** scheduling algorithm in xv6 that improves the efficiency of task execution
  - Enhanced xv6 memory management by integrating **lazy page allocation** to significantly improve memory utilization
- Understanding Linux Kernel Internals Through Custom Module Implementation — Spring'23
  Guide: Prof. Purushottam Kulkarni — Topics in Virtualization and Cloud Computing
  - IDesigned kernel modules to explore **kernel internals** having process listing and heap analysis functionalities
  - Enhanced modules to determine kernel stack pointers, map address spaces, and measure memory allocations
- 3D Visualization and Analysis of Seismic Volumes — Spring'23
  Guide: Prof. Prabhu Ramachandran — Parallel Scientific Computing and Visualization
  - Developed a visualization tool using the **Mayavi** and **TraitsUI** Python libraries for interactive geological analysis
  - Enhanced subsurface geological investigation through advanced geophysical analysis and multi-dimensional plotting
- Justice System and Prison Overflow — Spring'23
  Guide: Prof. Om P. Damani — System Dynamics: Modeling & Simulation for Development
  - Conducted a literature survey to identify factors contributing to prison overflow and developed a **system dynamics model** to simulate impact on prison population dynamics that provides insights for reforms to mitigate overcrowding
- Robust Mastermind Player — Spring'21
  Guide: Prof. Ashutosh Gupta — Logic for Computer Science
  - Formulated and implemented a player for the logic-based game Mastermind using **SAT** solving techniques and the Z3 Theorem Prover, which gives accurate performance even against adversary's inconsistent or unreliable feedback

Dept. of Computer Science, Virginia Tech

- Two-tier memory management for Compute Express Link (CXL) memory — Jul 2024 – Sep 2024
  Guide: Prof. Huaicheng Li — Remote
  - Integrated Data Access MONitor based memory management patches into the linux and reviewed the source code
  - Analyzed Redis performance on emulated CXL memory using **YCSB** benchmarks and compared results with vanilla linux memory management configurations to identify improvements and bottlenecks

TALKS

Compute and Communication trade-offs for scalable Large Language Models (LLMs)
Host: Prof. Purushottam Kulkarni, SynerG Lab, IIT Bombay — January 2024

AI-Infrastructure Reading Group, Microsoft Research India Lab
Flux: Fast Software-based Communication Overlap On GPUs Through Kernel Fusion — August 2024
Splitwise: Efficient generative LLM inference using phase splitting — April 2024

COURSE PROJECTS

Network Simulation — Spring'21
Implemented a File Transfer Protocol in **C** and analyzed throughput variations of TCP variants using Wireshark and NS3

Online Computing and Development Environment (IDE) (Code) — Fall'20
Developed a Django-based multi-language online IDE with real-time testing, file storage, and library/package support

Data Prefetchers and Cache Replacement Interaction (Code) — Fall'21
Compared cache replacement policies (LRU, Hawkeye) combined with prefetchers (PACMan, IPCP) across various traces

Multi-cycle RISC Processor (Code) — Spring'21
Implemented an 8-register, 16-bit multi-cycle processor with sync write and async read operations in VHDL

Real-Time Application Monitor — Spring'22
Developed an app to monitor system resources, with Telegraf for data collection and a time-series database for storage

KEY COURSEWORK

Systems and Networking:
Topics in Virtualization and Cloud Computing, Operating Systems, Computer Networks, Parallel Scientific Computing and Visualization, Database and Information Systems, Implementation of Programming Languages, Computer Architecture, Principal of Systems and Data Security, Digital Logic Design, Introduction to GPU Programming (Online)

AI/ML:
Introduction to AI/ML, Foundations of Reinforcement Learning, Automatic Speech Recognition

TECHNICAL SKILLS

Programming: CUDA, Python, C/C++, Java, MATLAB, Bash, SQL, Assembly
Software & Tools: PyTorch, LaTeX, Git, Lex, Yacc, Mayavi, TraitsUI, ChampSim, NS-3
Tools/Frameworks: HTML, CSS, JavaScript, Angular, Django

EXTRA -CURRICULAR ACTIVITIES

National Service Scheme (NSS), IIT Bombay — 2019 – 2020
Completed **80+** hours of community service at Social Development under the National Service Scheme
Associated with Parivartan, an initiative of the NSS, involving writing blogs on sustainable development

National Cadet Corps (NCC), Banaras Hindu University — 2015 – 2017
Awarded the National Cadet Corps (NCC) 'A; certificate for completing training in the Junior Division Air Wing
Attended the Annual Training Camp-311, NCC, which included rigorous physical training, drills, and sports

Interests: Hindi/Urdu Poetry — Present